Termproject

Introduction

This is a causal analysis on the properties on sale in Budapest. The data was scraped from ingatlan.com by me on the 16th of December, 2021. I investigate the effect of the size of the property on the price of the property.

I think analyzing housing prices is always meaningful, it is a crucial part of our daily lives, as everyone needs to live somewhere. People buy, sell and rent properties all the time. This is influenced by the prices of the properties. Understanding the change in the prices offers us the opportunity to have further investigations into what causes these changes.

My research question is simple: Are the houses on sale in Budapest on average more expensive if they are larger (in m^2)?

Data

The data is from ingatlan.com. It was scraped on the 16th of December, 2021. The scrapper went through every property listed under https://ingatlan.com/szukites/elado+lakas+budapest and all of the later pages. It scraped the address, the price, the number of rooms, the description, and the 20 element long list below the description. I have 27 908 observations and 27 variables. It is accessible on https://raw.githubusercontent.com/kanyipi/ingatlan/main/budapestall/budapestallfull.csv.

The first main variable is the price, how much the property costs. It is a double and it should be in millions of HUF. I converted all of the prices to millions of HUF with the help of price_in_cur which is either HUF or EUR. I converted the EUR ones to millions of HUF with live EUR to HUF rates. The other main variable is area, it is in m^2 it is format. Before cleaning it's format was " m,". I converted it to number and removed the string parts. The next variable I use is noroom which denotes the number of rooms and the number of half rooms. I calculate the final noroom to be number of full rooms + number of half rooms / 2, which makes the assumption that 2 half rooms are a normal room, which may or may not be true. The next variable is Ingatlan.állapota, which is the condition of the property I mapped them to integer values shown in appendix, where the higher the value the better the condition it is in. The last variable is Építés.éve which is the building year. The problem is that we have years after 2010 and intervals before 2010 in string. I mapped the intervals to their last year and tried to solve the issue by splining this variable at 2010.

The filtering is selecting these variables, the address, and the description, then removing the ones with NA values, and filtering on size and price, so that remaining ones are liveable houses.

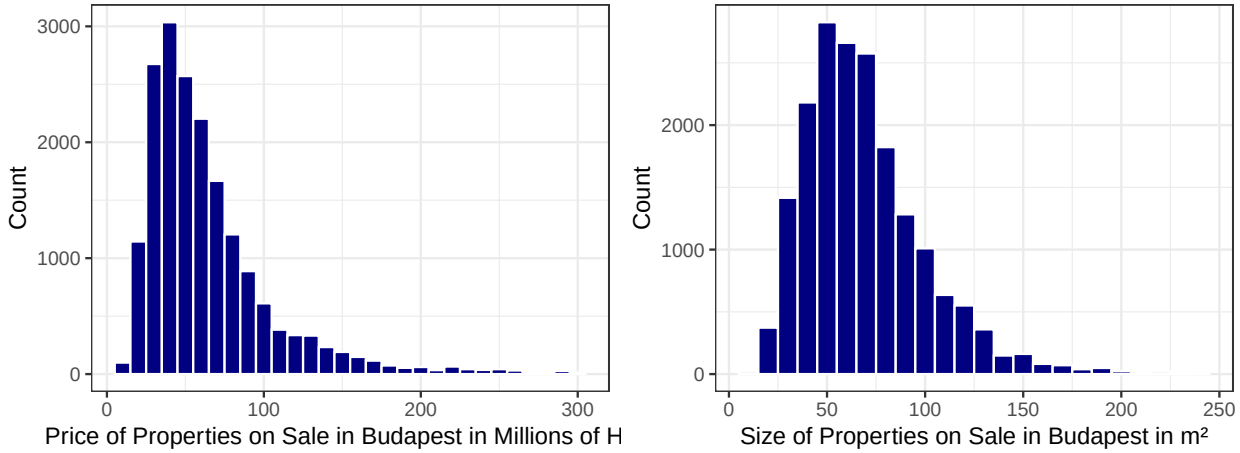
Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
Price <- price	65.12	54.50	41.68	6.90	299.99	23.50	147.50
Size <- area	69.60	64.00	31.10	11.00	240.00	30.00	127.00
'Number of Rooms' <- noroom	2.47	2.00	1.05	1.00	11.00	1.00	4.00
'Building Year' <- yearbuilt	1974.52	1980.00	27.03	1950.00	2025.00	1950.00	2022.00
Condition <- condition	3.45	3.00	1.36	1.00	6.00	1.00	6.00

The number of observations is 18343 for all of our key variables. These are after filtering on price so that it is greater than 5M HUF, but less then 300M HUF. Also filtered on size so it is less greater than 10 m^2, but less than 250 m^2. I choose these values as I feel like these are filtering for liveable houses. Usually bigger and more expensive properties were multiple houses, for investing. While smaller and cheaper properties were not really properties, but part of houses and people looking to swap houses.

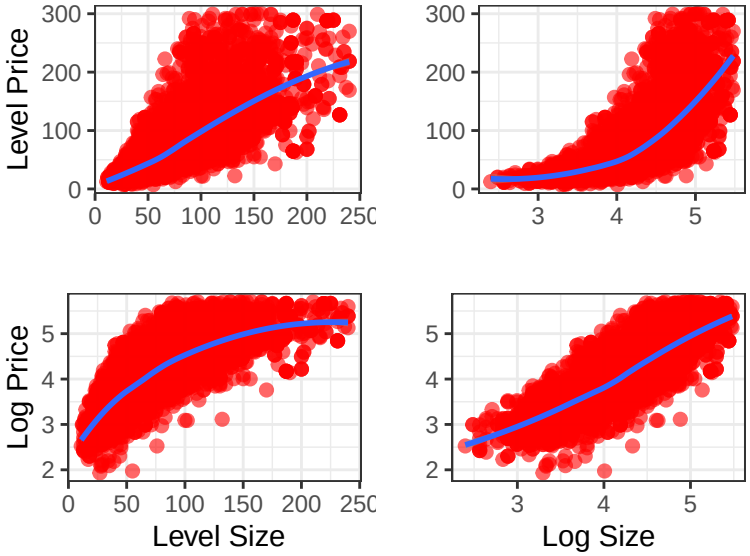
These are the summary stats of the already filtered dataset, so as we can see the minimums and maximum are close to the filter values. The mean of the price is 65.12M HUF the median is 54.5M HUF this means we have a long right tail. The mean of the Size is 69.60 m^2 the median is 64.0 m^2 this means we have a long right tail. The statistics for building year are not that meaningful because of the interval structure before 2010.

As the focus is on the price and on the size, the next Figure shows the histograms for these variables. I show the rest of the histograms in the appendix.



The long right tail on the histograms can be seen.

Here is the level, and the log associations of price and area:



The five models I made are :

- The first is: a simple linear regression of price on area both of them in logs
- The second is: area is splined with one knot at 4.5. This model is to see if the relationship is the same across the whole data
- The third is: add the building year as a control variable
- The forth is: number of rooms added as a control variable, this may not be that useful as number of rooms is maybe heavily correlated with the size
- The fifth is: condition added, year built splined at 2010

The choice of variables was mostly done by what interested me. There were some other interesting variables, for example on which floor the property is or what is the energy certificate of the property. For these I felt like these would be an additional "filter", as for example old houses don't have energy certificates and family houses don't have floors.

Model

My preferred model is:

$\log(\text{price}) = 0.15 + 0.99\ (log(size) < 4,5)\ 1.21\ (log(size) \geq 4.5) + \delta Z$

where Z are standing for the controls, which includes controlling for year built, number of rooms, and condition. From this model we can infer:

- alpa 0.15 is hard to interpret as both sides are in log, so we are more interested in the differences
- when the price of properties is 1 percent larger, but the log value of size is below 4.5, we see properties to be on average 0.99 percent more expensive.
- when the price of properties is 1 percent larger, but the log value of size is above or equal to 4.5, we see properties to be on average 1.21 percent more expensive.
- when the number of rooms is 1 unit larger, but the the number of rooms is below 4, we see properties to be on average 0.01 percent more expensive.
- when the number of rooms is 1 unit larger, but the the number of rooms is above or equal to 4, we see properties to be on average 0.04 percent less expensive.
- when the building year is 1 unit interval younger, but the the building year is below 2010, we see properties to be on average 0 percent less expensive.
- when the building year is 1 year younger, but the the building year is above or equal 2010, we see properties to be on average 0.01 percent more expensive.
- when the condition is 1 unit better, we see properties to be on average 0.1 percent more expensive.

Based on the heteroskedastic robust standard errors, these results are statistically different from zero. To show that, I have run a two-sided hypothesis test:

$$H_0 := \beta_1 = 0$$

$$H_A := \beta_1 \neq 0$$

I have the t-statistic as 94.58 and the p-value as 0, which confirms my conclusion.

I compare the models to learn about the stability of the parameters:

[H]

Table 2: Models to uncover relation between Size of Properties and Price of Properties

	(1)	(2)	(3)	(4)	(5)
Intercept	-0.3550***	-0.2318***	-6.804***	-6.462***	0.1494
	(0.0225)	(0.0276)	(0.1616)	(0.1757)	(0.2295)
Log Size	1.053***				
	(0.0055)				
Log Size (<4.5)		1.022***	1.021***	0.9625***	0.9850***
		(0.0069)	(0.0068)	(0.0112)	(0.0104)
Log Size (>=4.5)		1.181***	1.203***	1.208***	1.211***
		(0.0219)	(0.0220)	(0.0278)	(0.0249)
Building Year			0.0033***	0.0032***	
			(8.05e-5)	(8.36e-5)	
Number of Rooms < 4				0.0315***	0.0089*
				(0.0048)	(0.0044)
Number of Rooms >= 4				-0.0429***	-0.0391***
				(0.0087)	(0.0079)
Building Year <2010					-0.0003**
					(0.0001)
Building Year >=2010					0.0068***
					(0.0008)
Condition					0.0990***
					(0.0021)
Observations	18,343	18,343	18,343	18,343	18,343
R2	0.68353	0.68437	0.71052	0.71161	0.75667

Robustness Check, External Validity

I think the models are pretty robust as my question is is pretty easy, the R squares are starting from 0.68 to 0.75. This means that the 75% price variance is explained by the size variance. To see external validity, on Budapest on other dates, this could be done on a later date and reconfirmed. For other places we could scrape their websites, but I am confident that properties are usually more expensive when they are larger, at least for Budapest.

Conclusion

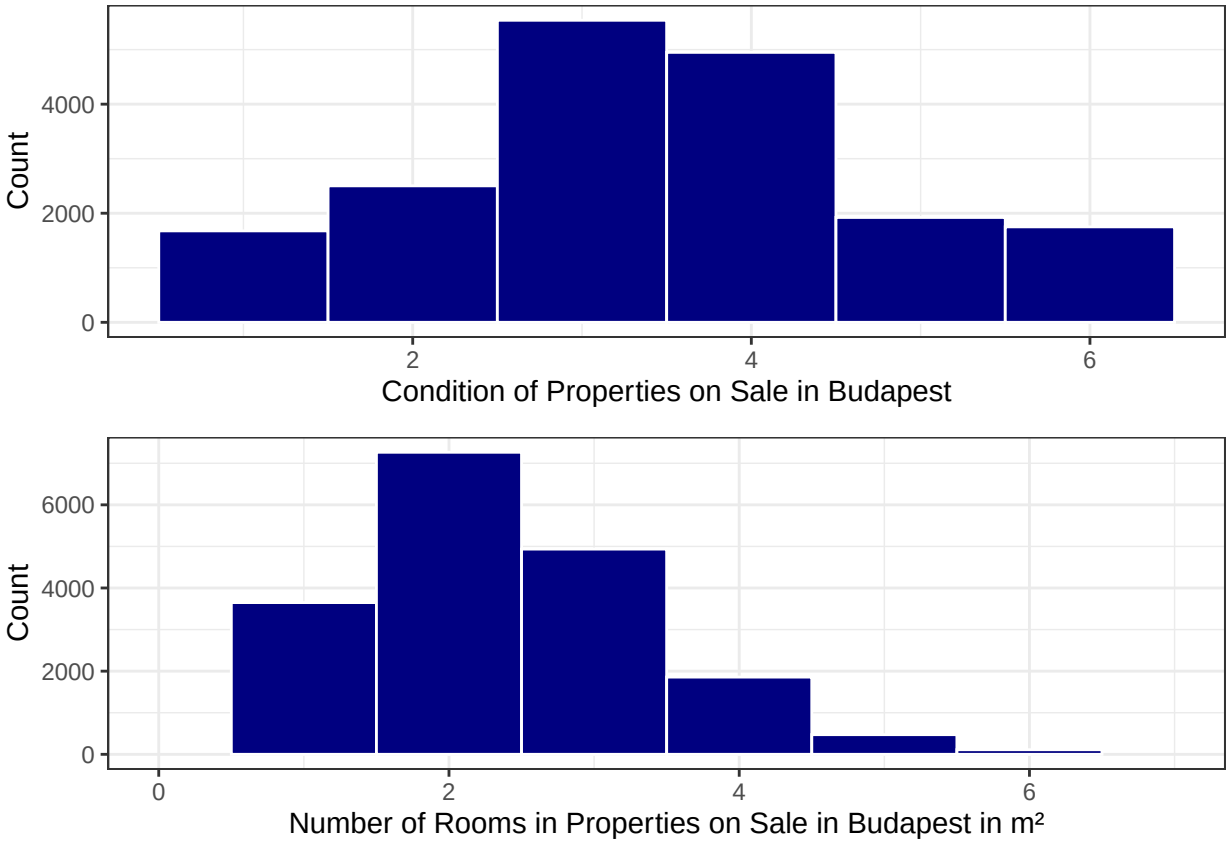
In conclusion I think properties are more expensive when they are larger at least in Budapest, I think this should be eternally valid in time well, a somewhat valid for other places too. There may be cultures where the smaller the property the better, or where people need to pay large sums of property tax on big properties making the condition more important. Some ideas for further work would be to scrape again and check the changes in prices. This would cause the problem of only the remaining properties being advertised. There could be also more variables included, but that may force us to only analyse some kind of properties. It is also noteworthy that out of the three categories of causality, I believe this is in the category where x causes the changes is y.

Appendix

Condition map

hungarian_condition	nincs megadva	befejezetlen	felújítandó	közepes állapotú	jó állapotú	felújított	újszeru	új építésu
english_condition	NA	not finished	needs renovation	mediocore condition	good condition	renovated	like new	newly built
integer_value	NA	1	2	3	4	5	6	7

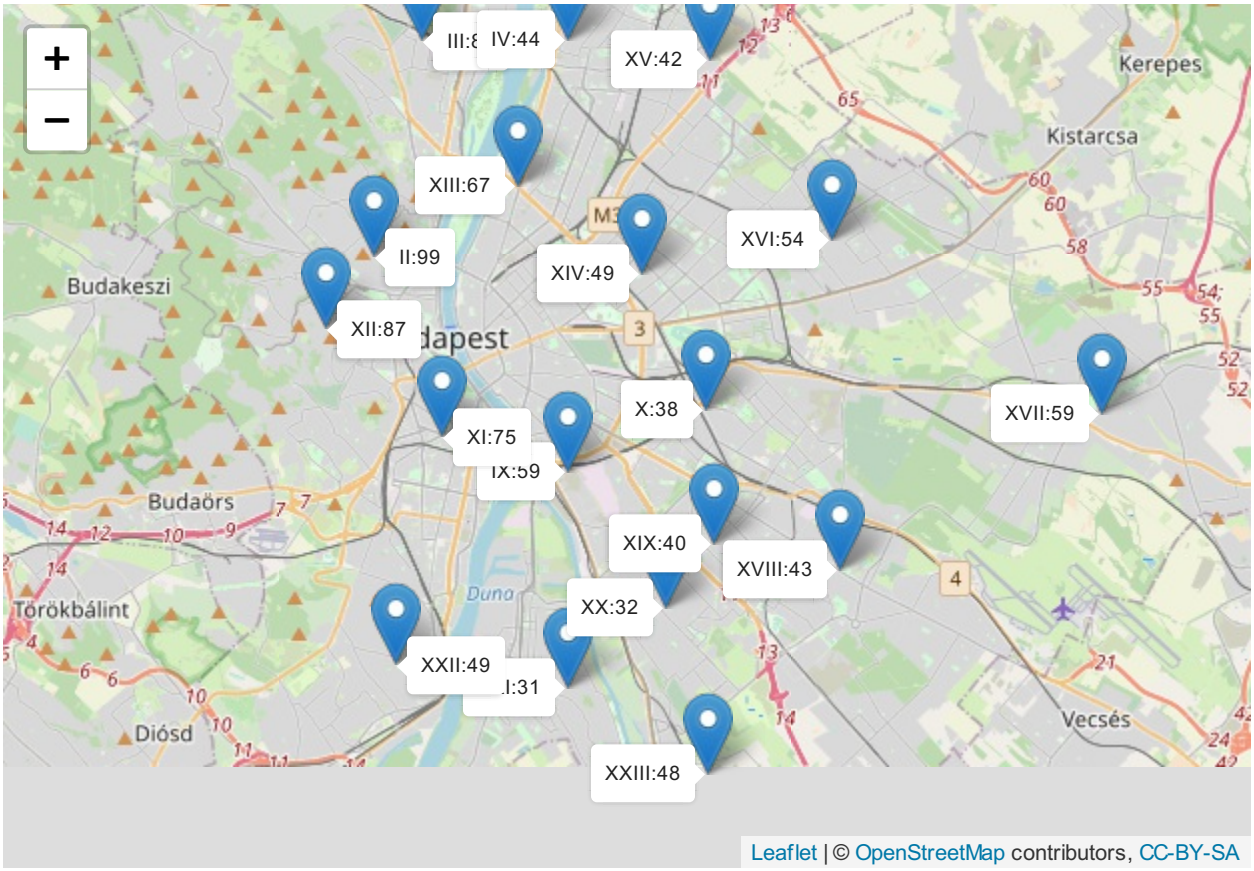
Histograms for Condition and Number of Rooms



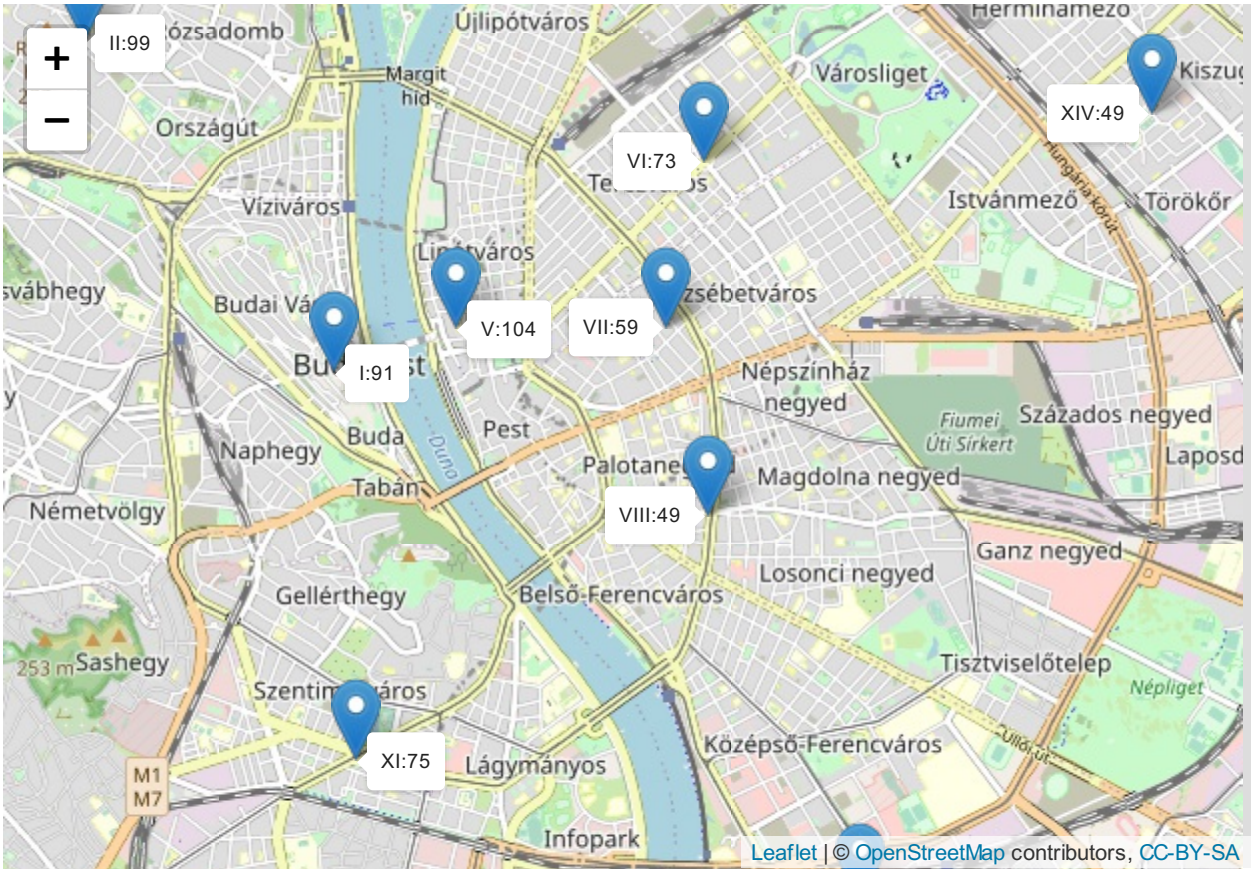
Some maps to visualise the mean price and size in each district

These maps were made by the leaflet package, it was hard to show them in a pdf report as leaflet is only knitable to html, but with the use of the webshot packages it is knitable to pdf. The coordinates of the districts are collected by hand from their hungarian wikipedia page. I merged the coordinate table to the average price and size for each district. Then I looped through them and added their labels to the map.

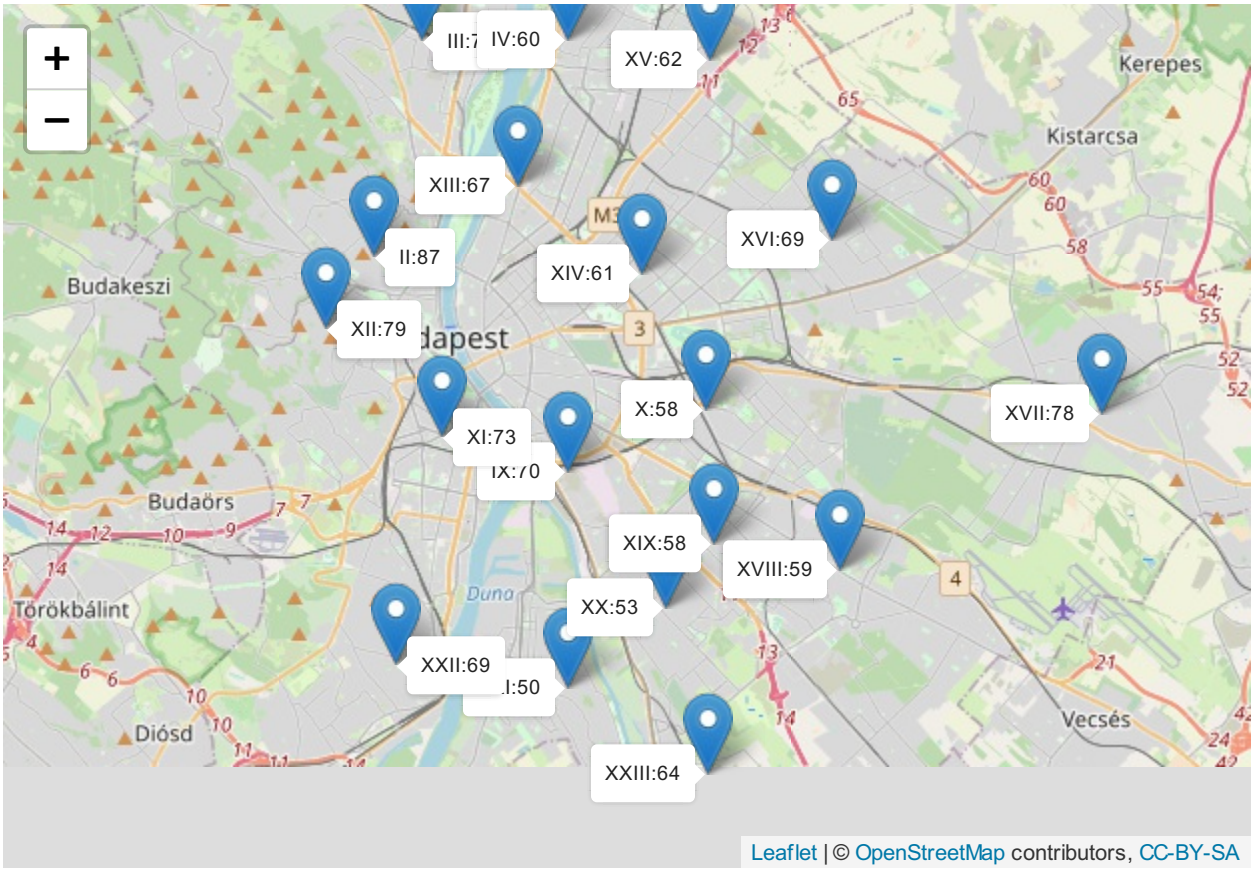
Average outer district prices of properties in millions of HUF for each district as label:



Average inner district prices of properties in millions of HUF for each district as label:



Average outer district sizes of properties in m^2 for each district as label:



Average inner district sizes of properties in m^2 for each district as label: